Predicting Degree Completion through Data Mining

Tatiana A. Cardona, Missouri University of Science and Technology

Tatiana A. Cardona is a Ph.D. candidate in Systems engineering at Missouri University of Science and Technology (MS&T)from where she also received her M.S. in Engineering Management in 2006. Tatiana completed her B.S. in Industrial Engineering at Technological University of Pereira, Colombia in 2009. from the same institution. Her research interests include statistical modeling, Operations research and Data Science. She has served as a head teaching assistant for four semesters in operations management and project management in the MS&T.

Dr. Elizabeth A. Cudney, Missouri University of Science & Technology

Dr. Elizabeth Cudney is an Associate Professor in the Engineering Management and Systems Engineering Department at Missouri University of Science and Technology. She received her B.S. in Industrial Engineering from North Carolina State University, Master of Engineering in Mechanical Engineering and MBA from the University of Hartford, and doctorate in Engineering Management from the University of Missouri – Rolla. In 2018, Dr. Cudney received the ASQ Crosby Medal for her book on Design for Six Sigma. Dr. Cudney received the 2018 IISE Fellow Award. She also received the 2017 Yoshio Kondo Academic Research Prize from the International Academy for Quality for sustained performance in exceptional published works. In 2014, Dr. Cudney was elected as an ASEM Fellow. In 2013, Dr. Cudney was elected as an ASQ Fellow. In 2010, Dr. Cudney was inducted into the International Academy for Quality. She received the 2008 ASQ A.V. Feigenbaum Medal and the 2006 SME Outstanding Young Manufacturing Engineering Award. She has published seven books and over 80 journal papers. Dr. Cudney is a certified Lean Six Sigma Master Black Belt. She holds eight ASQ certifications, which include ASQ Certified Quality Engineer, Manager of Quality/Operational Excellence, and Certified Six Sigma Black Belt, amongst others.

Dr. Jennifer Snyder, Valencia College

Jennifer Snyder serves as the Dean of Science for Valencia College's East Campus. She earned a Ph.D. from the Department of Engineering Management and Systems Engineering at Missouri University of Science and Technology. She received her B.S. and M.S. in Chemistry from Missouri State University in Springfield, Missouri.

Predicting Degree Completion through Data Mining

Abstract
Universities and colleges continuously strive to increase student retention and degree completion. The U.S. Department of Education has set the goal of preparing a society with individuals capable to "understand, explore and engage with the world" specific skills that can be achieved through STEM majors. Currently, considerable student data are collected and there is a latent opportunity to make the available information useful for determining the factors that influence retention and completion rates. Analyzing student data with those aims is vital for intentional student advising. To this end, this research presents the application of decision trees to predict degree completion within three years for STEM community college students. Decision trees also enable the identification of the factors that impact program completion using non-parametric models by classifying data using decision rules from the patterns learned. The model was developed using data on 283 students with 14 variables. The variables included age, gender, degree, and college GPA, among others. The results offer important insight into how to develop a more efficient and responsive system to support students.

Keywords: Student retention, decision trees, degree completion, engineering education

Introduction

One of the main concerns for universities and colleges is attrition rate. Students able to complete their degrees in the expected time directly impacts the reputation of the institution, as it reflects institutional commitment on contributing to the society by preparing individuals capable of engaging with the world (Williford & Schaller, 2005). Despite this, retention rates are currently low. With respect to college and university students pursuing STEM majors, retention rates are 69% and 48%, respectively (Snyder & Cudney, 2018).

Colleges and universities collect considerable student data. However, their ability to process the available information does not occur at the same pace as the collection (Morris, 2016). Therefore, effort needs to be made on making the data useful to improve student retention. For instance, by determining the factors that influence student retention and completion rates, it is possible to improve the intentional student advising, planning, and development of retention strategies based on student needs (Slim et al., 2005).

In recent years machine learning techniques have been applied to process educational data, which aligns with the focus on improving the processing of information. According to the literature, those techniques offer predictions of student dropout with high confidence (Pereira & Zambrano, 2017). Within machine learning techniques, decision trees (DT) have been employed successfully to predict and classify factors that impact student success measured as risk of dropout, attrition risk, and completion risk.

The purpose of this research was to develop a prediction model to forecast program completion within three years by STEM community college students and identify the factors that influence successful completion. To this end, this paper presents the application of DT as a machine learning technique using a data base comprised of 283 entries with 14 variables collected from a

community college in the Midwest. DT was used to develop a predictive model for student success. The key research question is: Can DT accurately predict student completion rates?

The remainder of this paper is structured into the following sections: literature review and background on DT applications on student success prediction, research methodology, results, and conclusions and future work.

Literature Review

DT have been one of the most frequently applied machine learning techniques for prediction of student success and identification of factors that influence it. According to Adejo and Connolly (2018), the advantage of DT resides on the computational speed and flexibility for modelling non-linearity. Further, DT structures are easy to understand and communicate; however, the main weakness is the overfitting/underfitting with an option to mild it by pruning. Several studies reflect the idea that DT offered a more visual structure of the results and state the importance of using the technique although other techniques could have better accuracy results (Delen, 2010; Delen, 2011; Oztekin, 2016). Research by Delen (2010, 2011) found that the classification of factors indicated that fall GPA, loans, and financial aid had a significant impact on predicting student attrition. Oztekin (2016) developed a hybrid method to predict completion for undergraduate students and also found that GPA was an important predictor variable.

Several studies applied principal component analysis (PCA) to a data set to filter the number of variables to be included in the model (Dissanayake et al., 2016; Adejo and Connolly, 2018). In the study by Dissanayake et al. (2016), not all techniques showed improvement in the results when applying PCA. Rather, DT showed better performance when using the original dataset.

In another study, Babić (2017) developed a classification model for predicting student academic motivation. The methodology included the application of machine learning classifiers such as neural network (NN), DT, and support vector machine (SVM). The results showed there was not a significant difference in the performance of the techniques. Supporting this conclusion Miranda and Guzman (2017) identified the factors that determine student dropout by applying different data mining techniques including Bayesian network classifier, DT, and NN. The results showed there was no significant difference within the performance of each technique.

Additional comparison of methods to identify key factors that impact the accuracy of an early-alert system was conducted to determine the level of factor importance. Pereira and Zambrano (2017) identified that the most relevant academic factors were low average in grades, number of failed classes in initial semesters, and department of study. Further, the relevant socioeconomic factors were university enrollment fee and provenance from south of the department. While, Tsao et al. (2017) concluded that the variables chosen for creating the datasets greatly impact the performance of the prediction models.

Uddin and Lee (2017) developed a hybrid model to predict a good fit in major for students to decrease dropout risk. Two algorithms that used several machine learning techniques including DT were integrated in the master algorithm to quantify the academic success factor. The results evidenced that the more data the more accurate the prediction. The hybrid method out-performed several known stand-alone techniques.

The DT methodology has been successfully used to predict academic success in higher education. However, the majority of the research has been performed in universities, rather than community colleges. The lack of research is this area indicates that more research should be performed to increase retention and completion of STEM students in community colleges.

Research Methodology

The data utilized for this research was collected from a community college located in Missouri. The community college offers associates degrees in STEM fields. Further, the community college allows students to declare their major upon entrance, which makes it ideal for data analysis. The data was collected over a five year period.

The research process was conducted in the following stages: 1) data description and preparation, 2) data modeling and application of DT, and 3) model assessment. A pictorial representation of the modeling process is provided in Figure 1. The stages are explained in more detail in the following subsections.

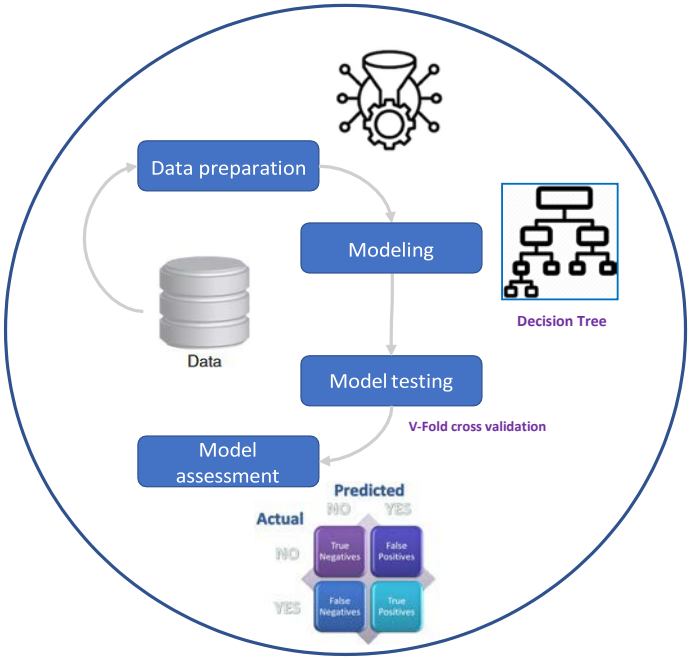


Figure 1. Data analytic methodology

Data description and preparation.

The data for this research was collected from a community college in the Midwest, which offers associate degrees in STEM majors. The dataset was comprised of five years of registered students, which consists of 904 students pursuing degrees in chemistry, biology, and engineering. From this data, 177 were identified as completing the degree within three years (150% of normal time for completion as required to be reported by the 1990 Student Right-to-Know Act for postsecondary institutions). The remaining 727 students did not graduate within that period, which is most commonly due to college withdrawal or switching to a non-STEM major. The data set was cleaned because of considerable missing and inconsistent data. For example, standardized exam scores were not available or provided for some students. After cleaning the data from incomplete records, a final dataset of 282 students was selected, which consisted of 51

completers and 231 non-completers. The data set had 14 variables, a non-exhaustive number for computational purposes. These variables were selected as they were readily collected and available. Therefore, it was not necessary to reduce the number of variables on the data. Table 1 provides a list of the variables used in the research.

Table 1. Variables used in the study

Variable	Type
Complete	Yes/No
Degree	Chemistry, Biology, Engineering
Age	Numerical
Gender	Female/Male
Full Time Student	Yes/No
1st Generation Student	Yes/No
Plan to work	Yes/No
ACT comprehensive	Numerical
ACT English	Numerical
ACT mathematics	Numerical
ACT reading	Numerical
High school GPA	Numerical
College GPA (Target variable)	Numerical

Data modeling

A DT is a tree like structure with a hierarchical nature. It can visually represent a decision-making process that divides the data as univariate splits for categorical predictor variables. The goal of DT is the prediction on a dependent variable, but also variable classification can be done by using this technique. The structure consists of classes (leaves), attributes (internal nodes), and connecting attributes (branches). It traces the path of nodes and branches to generate the prediction. DT are flexible in the fact that they examine the effects of the predictor variable one at time and can be computed for categorical and numerical predictors (Breiman et al., 1984).

In this study, classification and regression trees (CART) was used. This method for splitting selection generates an exhaustive search for univariate split producing the maximum goodness of fit. The stopping criteria selected was FACT. It allows for splitting until nodes contain no more cases than a specified fraction of the size of the class. For this study, 0.05 was the fraction used. It was also important to set the model to be equally precise for predicting students that could complete on time as for predicting the ones who could not. A cross validation of 10 folds was set in the training and a global cross validation was generated after running the training in order to validate the model. The model was implemented using Statsoft Statistica 12.

Model assessment

The model was assessed using measures of performance in training and the misclassification matrix. For testing the prediction, a 10-fold global cross validation was generated and the results were compared with the cross validation generated with the training. The overall performance is calculated as the proportion of correctly classified values from the sample size (N). For the identification of factors that impact the prediction, Statsoft Statistica 12 presents the results for predictor importance as a table with a ranking score in a range of 0-100 for each predictor.

Results

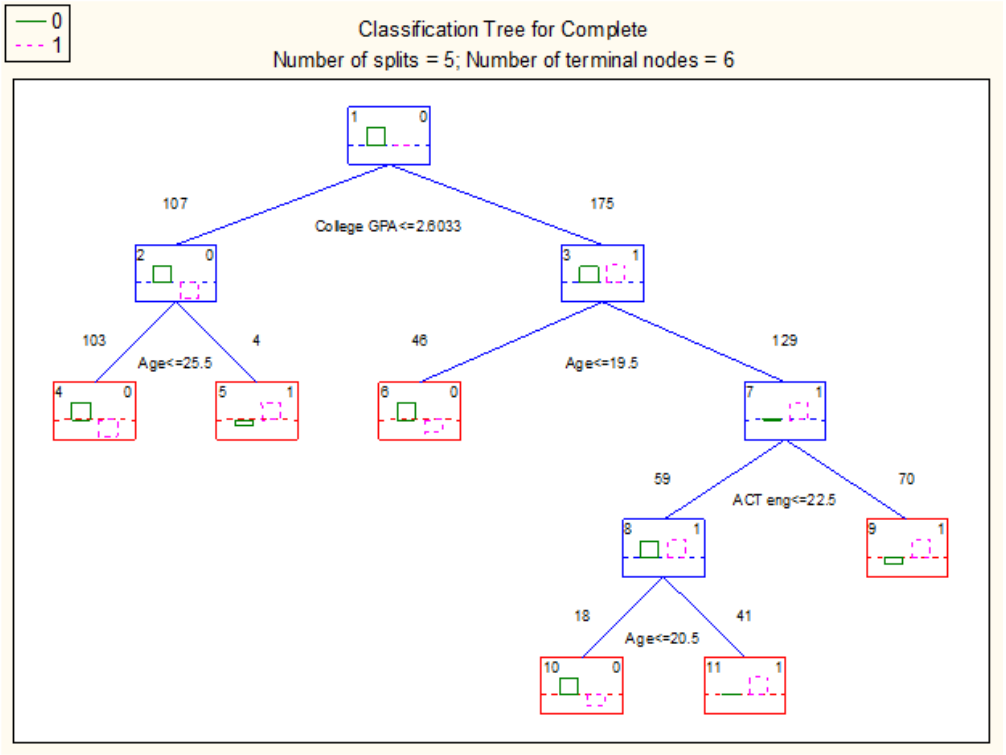
The selected tree had 11 nodes, within 6 are terminal nodes. The results are presented in Table 2 and Figure 2. Prediction class is 1 for completer or 0 for non-completer. Terminal nodes 4, 6, and 10 had a prediction of non-completer with 2, 5, and 3 misclassifications, respectively. While terminal nodes 5, 9, and 11 had prediction of completer with 1, 16, and 14 misclassifications, respectively. College GPA, age, and ACT Engineering were used as the splitting variable.

Table 2 Selected tree results

Tree Structure (subsample estratificado.sta)
Child nodes, observed class n's,
predicted class, and split condition for each node

Node	Left branch	Right branch	n in cls 0	n in cls 1	Predict. class	Split constant	Split variable
1	2	3	188	94	0	2.603295	College GPA
2	4	5	102	5	0	25.5	Age
3	6	7	86	89	1	19.5	Age
4			101	2	0	--	--
5			1	3	1	--	--
6			41	5	0	--	--
7	8	9	45	84	1	22.5	ACT eng
8	10	11	29	30	1	20.5	Age
9			16	54	1	--	--
10			15	3	0	--	--
11			14	27	1	--	--

Figure 2 Selected classification tree



The cost matrices from the training and test data are displayed in Table 3. The overall performance for the training and testing is consistent with not a significant difference (85.47% and

79.43%, respectively). The cross validation was also evaluated to ensure the consistency. Therefore, training cross validation cost and global cross validation cost and their respective standard deviations were compared for similarities (Table 4). In conclusion, the cost percentages in training and testing are very similar, which confirms consistency on the predictions.

Table 3 Misclassification matrix. Left, training data. Right, testing data.

Misclassification matrix Predicted (row) x Observed (column) Learning sample (N) = 282		
Class	0	1
0		10
1	31	

Global cross validation misclassification matrix Predicted (row) x Observed (column)		
Class	0	1
0		22
1	36	

Table 4. Results statistics. Left, training. Right, testing

Training tree statistics	
CV cost	0.1985
Std	0.0251

Test tree statistics	
CV cost	0.2057
Std	0.0241

The results indicate that the DT methodology offers a good prediction model for STEM degree completion for community college students with the specified variables with validation performance of approximately 80%.

After evaluating the prediction abilities of the model, it was important to identify the variables that impact the prediction. Table 5 and Figure 3 present the classification of level of importance of the different predictors. The results showed that the most significant variables are college GPA, age, ACT math, and ACT English.

Table 5 Predictor importance

Variable	Ranking
Gender	2
Full time student	19
Part time student	8
First generation	2
Plans to work	15
Degree	13
ACT Comprehensive	43
ACT English	48
ACT Mathematics	53
ACT Reading	31
High School GPA	43
College GPA	59
Age	100

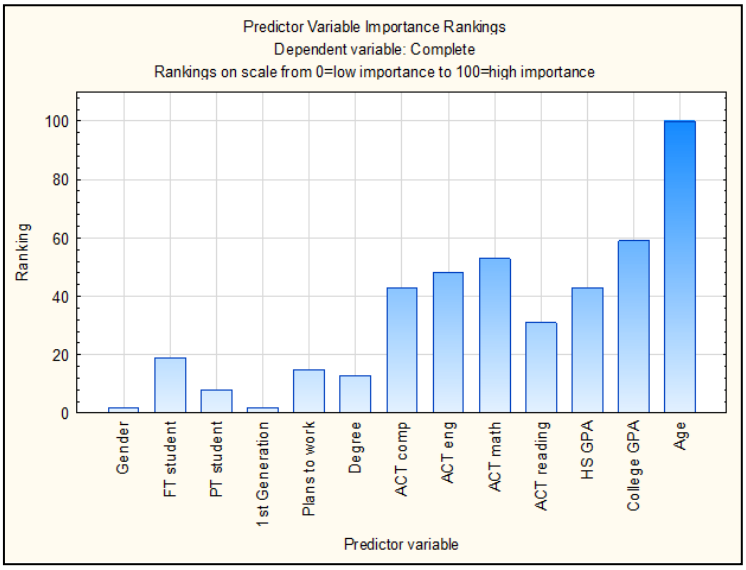


Figure 3 Predictor importance

Conclusions, limitations, and future work

This research presented a complete case of applying DT, which indicates that it is an effective tool for forecasting completion success of community college students in STEM majors. Also, it can be used for identifying the level of importance of the factors impacting such prediction. Although GPA is a common factor founded in prior literature as important for the prediction of student success, variables such as ACT math and ACT English are not commonly found in other studies. This statement infers what was found in the literature in terms of the variables chosen for the model impact its performance. Also, the findings suggest that the level of importance of those factors depended on the methodology used; however, further investigation should be performed.

As with any research study, there are limitations. First, the research findings are not generalizable as the study was conducted on data from only one community college. In addition, community colleges are representative of their local demographics. Therefore, results from one community college will not be generalizable to another university. However, the methodology should be applicable for the analysis. Next, the research was conducted using available data. The community college had information only on 14 variables. Numerous additional variables were identified through the literature. Future research should utilize data collected using considerably more data as noted in the relevant literature.

Further studies can also focus on combining a more complete mixture of factors to have a more robust model. In that manner a prediction model with the right set of variables can represent a useful tool for the creation of retention strategies by addressing the advising.

References

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education, 10*(1), 61-75.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*. *8*, 443-461.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees, Chapman and Hall/CRC Press, Boca Raton, Florida.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17-35.

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). Predictive Modeling for Student Retention at St. Cloud State University. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación universitaria*, *10*(3), 61-68.

Morris, L. V. (2016). Mining Data for Student Success. *Innovative Higher Education*, *41*(3), 183-185.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, *116*(8), 1678-1699.

Pereira, R. T., & Zambrano, J. C. (2017, December). Application of Decision Trees for Detection of Student Dropout Profiles. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (pp. 528-531). IEEE.

Slim, A., Heileman, G. L., Kozlick, J., & Abdallah, C. T. (2014, December). Predicting student success based on prior performance. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on* (pp. 410-415). IEEE.

Snyder, J., & Cudney, E. A. (2018, June), *A Retention Model for Community College STEM Students* Paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. https://peer.asee.org/29719

Tsao, N. L., Kuo, C. H., Guo, T. L., & Sun, T. J. (2017, July). Data Consideration for At-Risk Students Early Alert. In *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on* (pp. 208-211). IEEE.

Uddin, M. F., & Lee, J. (2017). Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction. *Social Network Analysis and Mining*, *7*(1), 29.

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In *Proceedings of the 45th Annual Forum of the Association for Institutional Research*